DNA Sequence Evaluation

# DNA Sequence Evaluation

Recall:

- **DNA molecules** and **DNA molecule fragments** are represented as **strings of letters** in an alphabet of **nucleotides**: $\{A, T, C, G\}$.
- Triples of nucleotides, called **codons**, encode one of **20 amino acids**.
- **DNA molecules** consist of **coding regions**, otherwise known as **genes** and **non-coding regions**, somtimes referred to as **junk DNA**. Codons in the **coding regions** regions are *transcribed* into **RNA** molecules, which, in turn are *translated* into proteins. The amino acids represented by the codons found in the genes form the *building blocks* for protein construction.
- Codons in **non-coding regions** are **not used** for protein production. However, some parts of non-coding regions may play other important biological roles associated with how the process of DNA transcription actually takes place.
- A **DNA molecule** consists of two strands, commonly referred to as *top* and *bottom*; *positive* and *negative*; or *normal* and *reverse*. The contents of one strand form a **reverse complement** of the other strand. The "complement" part: `A`(denine) is always paired with `T`hymine, while `C`ytosine is always paired with `G`uanine. The "reverse" part: the strand are read in opposite directions: the `5'` end of one strand corresponds to the `3'` end of the other strand and vice versa.
- Coding regions can appear on either DNA strand.

# GC Content

We add to the background knowledge above one more piece of biochemical information:

- There is a difference between the AT and the CG pairs of nucleotides in DNA.

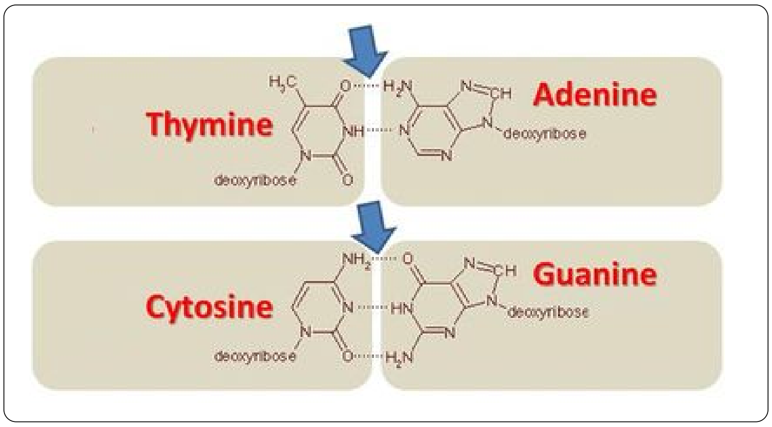


Figure 1: Hydrogen bonds between AT and CG pairs.

- Adenine and Thymine are connected by **two hydrogen bonds**.
- Cytosine and Guanine are connected by **three hydrogen bonds**. Figure 1 illustrates the chemical structure of the AT and CG bonds[1].
- It is assumed that the GC bond is stronger than the AT bond.
- The stronger the bond, the harder it is to break the two DNA strands apart.
- The easier it is to break two DNA strands apart, the *more chemically active* the DNA molecule (or its specific part) will be.
- Because GC bond is considered stronger, we expect that the difficulty of splitting DNA strands will increase with the increase in the number of GC base pairs in a DNA string.

GC Content quantifies that.

**Definition.** Given a DNA string $d$, its GC content, a.k.a., its textsfGC percentage is the percent of `GC base pairs` in $d$.

**Example.** Consider the following DNA string:

```
ATATGTAACT
```

In this string, eight out of 10 characters are `AT` and two are `CG`. The GC-content of this string is therefore 20%.

**Proposition.** GC content of a DNA string $d$ is the same as the GC content of $d$'s reverse complement.

### Computing GC-content

Tasks:

1. Given a DNA sequence in nucleotide alphabet, compute its GC-content.

[1] `http://en.wikipedia.org/wiki/File:AT-GC.jpg`

| Amino Acid | Three-letter code | One-letter code | Degeneration Level | Codons |
|---|---|---|---|---|
| Alanine | *Ala* | A | 4 | GCT, GCC, GCA, GCG |
| Arginine | *Arg* | R | 6 | CGT, CGC, CGA, CGG, AGA, AGG |
| Asparagine | *Asn* | N | 2 | AAT, AAC |
| Aspartic acid | *Asp* | D | 2 | GAT, GAC |
| Cysteine | *Cys* | C | 2 | TGT, TGC |
| Glutamic acid | *Glu* | E | 2 | GAA, GAG |
| Glutamine | *Gln* | Q | 2 | CAA, CAG |
| Glycine | *Gly* | G | 4 | GGT, GGC, GGA, GGG |
| Histidine | *His* | H | 2 | CAT, CAC |
| Isoleucine | *Ile* | I | 3 | ATC, ATT, ATA |
| Leucine | *Leu* | L | 6 | TTA, TTG, CTT, CTC, CTA, CTG |
| Lysine | *Lys* | K | 2 | AAA, AAG |
| Methionine | *Met* | M | 1 | ATG |
| Phenylalanine | *Phe* | F | 2 | TTT, TTC |
| Proline | *Pro* | P | 4 | CCT, CCC, CCA, CCG |
| Serine | *Ser* | S | 6 | TCT, TCC, TCA, TCG, AGT, AGC |
| Threonine | *Thr* | T | 4 | ACT, ACC, ACA, ACG |
| Tryptophan | *Trp* | W | 1 | TGG |
| Tyrosine | *Tyr* | Y | 2 | TAT, TAC |
| Valine | *Val* | V | 4 | GTT, GTC, GTA, GTG |

Table 1: Amino acids and their degeneration level.

2. Given a DNA sequence in nucleotide alphabet and a substring in the sequence, compute the GC-content of the substrng provided.

3. Given a DNA sequence in nucleotide alphabet (or a full chromosome) create a graph of GC-content variation throughout the given sequence.

## Codon Usage Bias

**Genetic code is degenerate.** There are $4^3 = 64$ possible codons. At the same time, there are only **20** amino acids; 21 if the **stop codons** are considered.

The **genetic code** is a mapping from 64 combinations onto 21 possible "outcomes". Biologists call such mappings **degenerate**.

**k-fold degenerate amino acids.** An amino acid is said to be **k-fold degenerate** if there are exactly $k$ different codons that map to it in the genetic code.

**Example.** Phenylanine (F) is **2-fold degenerate** as it is represented by two different codons, TTT and TTC in the genetic code. Similarly, Arginine (R) is **6-fold degenerate**: it is represented by six different codons: CGT, CGC, CGA, CGG, AGA and AGG.

Table 1 specifies the degeneration level for each amino acid. Table 2 shows how many $k$-fold degenerate amino acids there are for each $k$ from 1 to 6 (with 6 being the largest level of amino acid degeneration in the genetic code).

| Degeneration level | Number of amino acids | Amino acids |
|---|---|---|
| 1-fold | 2 | M, W |
| 2-fold | 9 | N, D, C, E, Q, H, K, F, Y |
| 3-fold | 1 | I |
| 4-fold | 5 | A, G, P, T, V |
| 5-fold | 0 | |
| 6-fold | 3 | R, L, S |

Table 2: Summary of degeneration levels in the genetic code.

**Synonymous codons:** codons that code for the same amino acid. E.g., GCT, GCC, GCA and GCG are all synonymous, as they all code for Alanine.

**Codon usage bias.** *Codon usage bias* of a DNA sequence is the difference (distribution) of the occurrences of *synonymous codons* in the sequence.

**Why codon usage bias.** It has been found that different organisms (and types of organisms) have *different codon usage bias*, i.e., some types of organisms (e.g., bacteria) tend to favor some of the codons to represent specific amino acids, while other organisms (e.g., mammals) tend to use different codons to represent the same amino acids.

Biologists believe that by studying the codon usage bias of a specific DNA fragment and comparing it to the codon usage bias of known DNA, they can discover similarities and dissimilarities in the DNA purpose, as well as understand specific property of some of the DNA regions[2].

## Measures of Codon Usage Bias

In practice **Codon usage bias** is quantified using a number of different measures. Some of the more popular ones are discussed here.

### Histograms

Not really a measure per se, but the *histogram representation* of the codon usage bias essentially underlies every other measure.

**Histogram.** A **histogram** is a mapping between a set of items and the number of times each item occurs in a given collection of items, and/or a mapping between a set of items and the percent of occurrence of each item in a given collection of items.

A histogram of the codon usage bias for a single amino acid in a DNA sequence $d$ is the mapping between the codons that code for the amino acid and the number of times each occurs in $d$.

**For example**, consider a DNA sequence (spaces are added for clarity) $d =$GCT GCT GCA GCG GCA GCA GCG GCT representing (in the first reading frame)

[2]For example, bacteria have significantly higher rates of protein synthesis than, e.g., individual cells in human body. This suggests to biologists that the codons used primarily in bacterial DNA facilitate high rate of DNA transcription/reproduction, while the codons used primarily in human DNA don't.

the amino acid sequence AAAAAAAA (eight Alanines). The codon usage bias histogram for this sequence is (using both the numeric count and the percent of occurrences):

| Codon | Count | Percent |
|---|---|---|
| GCT | 3 | 37.5% |
| GCC | 0 | 0.0% |
| GCA | 3 | 37.5% |
| GCG | 2 | 25.0% |

**Frequency of Optimal Codons**

**Optimal codon.** Biologists determined empirically a set of optimal codons in the genetic code, based on the frequency of their occurrence in a number of bacterial genomes.

Essentially, a codon is **optimal** if it occurs more frequently in the *high-expression genes* (i.e., genes that are used for protein synthesis often) than in *low-expression genes* (i.e., genes that are used for protein synthesis less frequently).

**Frequency of optimal codons measure.** Given a DNA sequence $d$ and a list $L$ of optimal codons, we break total number of codons in $d$ into three components:

- $X_{op}$: total number of occurrences of *optimal codons* in $d$.
- $X_{ex}$: total number of occurrences of *excluded codons* in $d$.
- $X_{non}$: total number of occurrences of *non-optimal codons* in $d$.

**Excluded** codons are **stop** codons, as well as codons for Methionine and Tryptophan (1-fold degenerate amino acids), and the codons for any other amino acids, for which an *optimal* codon is not supplied.

**Non-optimal** codons are all codons that are not optimal and not excluded.

The **frequency of optimal codons** measure, denoted $F_{op}$ is defined as

$$F_{op} = \frac{X_{op}}{X_{op} + X_{non}}.$$

**Which codons are optimal?** There is no predetermined set of optimal codons. Using different organisms are reference points will yield different sets of optimal codons. For example, Moriyama suggests that for the set of the five 4-fold degenerate amino acids, codons CCG, ACC, GTT, GCG and GGT may be considered optimal for *E.coli* bacterial genome, while for *Saccharomyces cerevisiae* bacterial genome, the CCA, ACT, GTT, GCT and GGT are optimal. These two sets differ on three codons.

**Implementation notes.** Since there isn't a single set of optimal codons, the implementation of this measure must take the set of optimal codons as an input parameter. Upon receiving the input: the list of optimal codons and the DNA string, the computation shall proceed as follows:

1. Determine *excluded* and *non-optimal* codons.
2. Compute the number of occurrences of each codon in the input string.
3. Aggregate the number of occurrences for optimal ($X_{op}$) and non-optimal ($X_{non}$) codons.
4. Compute the **frequency of optimal codons** measure as specified in the formula above.

#### Relative Synonymous Codon Usage (RSCU)

**Idea.** The **RCSU** measure is applied to an individual codon. It represents the ratio between the observed number of occurrences of a given codon and the its expected number of occurrences under the uniform distribution assumption.

**Formula.** Let $d$ be a DNA sequence. Let $a$ be an $n$-fold degenerate amino acid and let $i \leq n$ be its $i$th codon. Let $X_i$ be the number of occurrences of the codon $i$ in $d$. Then, $RCSU_i$ is defined as follows:

$$\mathbf{RCSU}_i = \frac{X_i}{\frac{1}{n}\sum_{j=1}^{n} X_j}$$

**Implementation notes.** **RSCU** numbers are used in some of the other measures discussed below. The input of **RSCU** computation is the DNA string (as everywhere else) and an amino acid. The output is the list of **RSCU** values for each codon representing this amino acid. Alternatively, you can supply a codon as input to the **RSCU** computation, but internally, you still will need to find what amino acid it is, and build the histogram for this amino acid (or use an already prepared histogram).

#### Codon Adaptation Index (CAI)

**Idea.** **CAI** (Codon Adaptation Index) estimates the extent of bias toward codons that occur more frequently in the DNA sequence.

**Preliminaries.** A **relative adeptedness** value $w_i$ for codon $i$ of an amino acid $a$ is computed as

$$w_i = \frac{\mathbf{RCSU_i}}{\mathbf{RCSU_{max}}} = \frac{X_i}{X_{max}},$$

where $X_{max}$ and $\mathbf{RCSU_{max}}$ represent, respectively, the number of occurrences and the **RCSU** of the *most frequent codon* for $a$ in a given DNA string $d$.

Let $g = a_1 \ldots a_n$ be a gene, and let it contain $L = n/3$ codons $c_1, \ldots c_L$.

**Codon Adaptation Index** of the gene $g$ is defined as follows:

$$CAI = \frac{CAI_{obs}}{CAI_{max}},$$

where

$$CAI_{obs} = \left( \prod_{i=1}^{L} RSCU_i \right)^{\frac{1}{L}},$$

where $RSCU_i$ is the $RSCU$ of the codon $c_i$ in the genome, and

$$CAI_{max} = \left( \prod_{i=1}^{L} RSCU_{i_{max}} \right)^{\frac{1}{L}},$$

where $RSCU_{i_{max}}$ is the largest $RSCU$ of any codon synonymous to $c_i$.

(Note, occurrences of $M$ and $W$ - the two non-degenerate amino acids, as well as stop codons are ignored in this computation).

Alternatively, one can express $CAI$ as follows:

$$CAI = \left( \prod_{i=1}^{L} w_i \right)^{\frac{1}{L}}$$

**Effective Number of Codons**

**Idea.** Need a measure that is not based on optimal codons.

**Preliminaries.** Consider a DNA sequence $d$. Given a $k$-fold degenerate amino acid $a$, let $n_{a1}, \ldots n_{ak}$ be respectively the numbers of occurrence of each codon coding for $a$ in $d$ (such that $\sum_{j=1}^{k} n_{aj} = n_a$ is the total number of occurrences of $a$ in $d$). We define the quantity $S_a$ as follows:

$$S_a = \sum_{j=1}^{k} \left( \frac{n_{aj}}{n_a} \right)^2.$$

Using $S_a$, we define $F_a$ as

$$F_a = \frac{n_a \cdot S_a - 1}{n_a - 1}.$$

Given $k \in \{2, 3, 4, 6\}$, $F_k$ is defined as the average value of $F_a$ for over all $k$-fold degenerate amino acids $a$.

**Formula.** Using the quantities established above, the **effective number of codons**, denoted as $N_c$ is computed as follows:

$$N_c = 2 + \frac{9}{F_2} + \frac{1}{F_3} + \frac{5}{F_4} + \frac{3}{F_6}.$$

**Explanation.** $N_c$ computes *the number of (distinct) equally used codons that would generate the same codon usage bias as observed.*

**Range.** $N_c$ range:

- **from 20**: for the strongest bias (only one codon per amino acid used);
- **to 61**: no bias — all possible codons (excluding stop codons) are used equally;